How AI can distort human beliefs

Year: 2023 | Citations: 114 | Authors: Celeste Kidd, Abeba Birhane

Abstract

Models can convey biases and false information to users Individual humans form their beliefs by sampling a small subset of the available data in the world. Once those beliefs are formed with high certainty, they can become stubborn to revise. Fabrication and bias in generative artificial intelligence (AI) models are established phenomena that can occur as part of regular system use, in the absence of any malevolent forces seeking to push bias or disinformation. However, transmission of false information and bias from these models to people has been prominently absent from the discourse. Overhyped, unrealistic, and exaggerated capabilities permeate how generative AI models are presented, which contributes to the popular misconception that these models exceed human-level reasoning and exacerbates the risk of transmission of false information and negative stereotypes to people.